Anesthesiology
1996; 85:1235-45
© 1996 American Society of Anesthesiologists, Inc.
Lippincott-Raven Publishers

Statistical Modeling to Predict Elective Surgery Time

Comparison with a Computer Scheduling System and Surgeon-provided Estimates

Ian H. Wright, M.B., B.S.,* Charles Kooperberg, Ph.D.,† Barbara A. Bonar, M.S.,‡ Gerard Bashein, M.D., Ph.D.§

Background: **Accurate estimation of operating times is a prerequisite for the efficient scheduling of the operating suite. The authors, in this study, sought to compare surgeons' time estimates for elective cases with those of commercial scheduling software, and to ascertain whether improvements could be made by regression modeling.**

Methods: **The study was conducted at the University of Washington Medical Center in three phases. Phase 1 retrospectively reviewed surgeons' time estimates and the scheduling system's estimates throughout 1 yr. In phase 2, data were collected prospectively from participating surgeons by means of a data entry form completed at the time of scheduling elective cases. Data included the procedure code, estimated operating time, estimated case difficulty, and potential factors that might affect the duration. In phase 3, identical data were collected from five selected surgeons by personal interview.**

Results: **In Phase 1, 26 of 43 surgeons provided significantly better estimates than did the scheduling system ($P < 0.01$), and no surgeon was significantly worse, although the absolute errors were large (34% of 157 min average case length). In phase 2, modeling improved the accuracy of the surgeons' estimates by 11.5%, compared with the scheduling system. In phase 3, applying the model from phase 2 improved the accuracy of the surgeons' estimates by 18.2%.**

Conclusions: **Surgeons provide more accurate time estimates than does the scheduling software as it is used in our institution. Regression modeling effects modest improvements in accuracy. Further improvements would be likely if the hospital information system could provide timely historical data and feedback to the surgeons. (Key words: Operating room management: statistical modeling; surgical scheduling.)**

This article is accompanied by an editorial. Please see: Dexter F, Macario A: Applications of information systems to operating room scheduling. ANESTHESIOLOGY 1996; 85:1232-4.

COST containment is, currently, a major priority in health care. Hospital surgical suites are likely targets for cost reduction efforts because they can consume 9% of an institution's annual budget.[1] Inaccurate scheduling of elective operations can increase costs, either when overestimation of operating time results in unused operating rooms or when underestimation results in unplanned overtime work or cancellation of cases. In studies, researchers found operating room utilization rates (operating time as a proportion of a notional 8-h working day) to be low overall, approximately 40-60% in widely varying settings, including the Chicago area,[2] the nation of Colombia,[3] and the United States Department of Veterans Affairs.[2] In addition, low utilization may occur despite frequent overrunning of the elective schedule. For example, in a British study of 3,657, half-day (3.5-h) operating schedules, it was found that 32.5% lasted less than 2.5 h (*i.e.*, less than 71.4% used), whereas 20.9% ran overtime by a half hour or more.[4] Inaccurate surgical scheduling also can have an economic impact on other aspects of hospital operation, including the recovery room,[5] intensive care unit, and ancillary services such as x-ray and clinical laboratories.

Despite the obvious importance of accurate scheduling in surgical suites, the subject has generated little in the way of scientific inquiry, in contrast to the extensive research on job shop scheduling in industry.[6] In the few published reviews on surgical scheduling[7,8] researchers

* Assistant Professor of Anesthesiology.

† Assistant Professor of Statistics.

‡ Graduate Student in Statistics.

§ Professor of Anesthesiology.

Received from the Departments of Anesthesiology and Statistics, University of Washington, Seattle, Washington. Submitted for publication October 24, 1995. Accepted for publication July 1, 1996. Supported in part by National Science Foundation Grant BCS-9315497 (G.B.). Presented in part at the annual meeting of the American Society of Anesthesiologists, San Francisco, California, October 15-19, 1994. Study conducted at the University of Washington Medical Center, Seattle, Washington.

Address reprint requests to Dr. Wright: University of Washington, Department of Anesthesiology, Box 356540, Seattle, Washington 98195-6540. Address electronic mail to: iwright@u.washington.edu.

discussed articles with limited scope, no practical assessment, or only short assessment periods. More recent reports also contained little in the way of objective data.[9–11] Although no particular method of scheduling has been shown to be superior, many institutions use a variant of "block scheduling," in which blocks of time are reserved for particular surgeons until some deadline, after which the remaining time becomes generally available.

Regardless of the method used to construct the daily schedule, having an accurate prediction of operating time for each surgery is a prerequisite for matching workload to capacity. Many institutions use commercial scheduling software to generate time estimates, to track the bookings, and to print each day's schedule. Our institution's scheduling software uses only historical data to make its estimates and does not take into account surgeon input or patient-related factors. Often, gross discrepancies occur, in either direction, between the scheduled and actual operating time. We hypothesized that the accuracy of operating time estimation could be improved if the surgeon's sophisticated knowledge about the degree of complexity of each individual patient were incorporated into a regression model and that any biases in individual surgeon's time estimates could be compensated for by modeling. To test these hypotheses, a study was designed in three phases to compare the accuracy of the scheduling software, the surgeon, and computer model-generated estimates of operating time for elective cases.

Materials and Methods

The study was performed at the University of Washington Medical Center (UWMC), a 450-bed adult tertiary referral center. Institutional approval was obtained for review of the medical records and enrollment of surgeons as subjects in the study. The operating time estimates and the daily operative schedule are generated by Surgiserver software (Serving Software, Minneapolis, MN), running under MS-DOS (Microsoft, Redmond, WA) on a network of IBM-compatible computers. The method for calculating time estimates is to take the actual duration (*i.e.*, from incision until the dressing is applied, as recorded by the operating room nurse at the time of surgery) of the last ten similar cases for an individual surgeon (grouped by an in-house coding system of approximately 1,200 descriptors) and calculate a trimmed mean by discarding the greatest and least durations and averaging the remaining eight. When fewer than ten cases are available, a simple mean is calculated. If there are no previous similar cases, an estimate is given by the head nurse of the surgical suite. This was a frequent occurrence when the database was first used, but now occurs only rarely, because the in-house coding system has sufficiently broad categories to include most surgeries. To improve the accuracy of subsequent case estimates, all cases are recorded using the in-house code for the operation actually performed, rather than the one originally scheduled.

Phase 1: Retrospective Review

The first phase of the study was a retrospective review of calendar 1993 data, comparing the time estimates generated by the computer scheduling system and those provided by surgeons with the actual duration of each case. The review was limited to the 43 surgeons who provided time estimates for 50 or more cases during the year. They were aware that their estimates were not actually used in the scheduling process.

The surgeons' estimates were grouped in 30-min-wide bins (0–29 min, 30–59 min, etc.) by actual operating time. For each case, the squared error between the predicted and actual operating times was calculated. Within each 30-min bin, these squared errors were then normalized using the average and the standard deviation of the squared errors of all the procedures within that bin, yielding a normalized squared error for each case. A score for each surgeon was calculated as the mean for that surgeon's cases from all bins. Under the null hypothesis that all surgeons are equally good predictors, the scores are approximately normally distributed because of the central limit theorem, even if the original prediction errors are not normally distributed. Therefore, the expected score and standard deviation could be calculated using the average and standard deviation of the individual scores. A similar calculation was performed on the scheduling system estimates. Using these scores, individual surgeon's estimates were compared with other surgeons' estimates and also with scheduling system estimates using the appropriate standard errors and normal tables.

Phase 2: Prospective Data Collection Using a Structured Form

In the second phase of the study (calendar 1994), data were collected prospectively from 64 surgeons who, after an explanation of the purpose of the study and the surgical input sought, indicated a willingness to participate and gave written informed consent. Most surgical services were sampled, although cardiac surgery estimates were not collected because the majority of their operations were classified as urgent, therefore falling outside the ambit of the study (cases appearing on the printed schedule compiled the day before surgery). The neurosurgeons initially did not participate in the study, although one neurosurgeon agreed to participate during phase 3. At the beginning of phase 2 and every 3 months during it, the surgeons were given a list of individual case and mean operating times for the previous 3 months (historical data), grouped by (up to three) surgical CPT codes (the American Medical Association's standardized Current Procedural Terminology coding system). To provide reports with CPT codes, mappings from the Surgiserver data base to the Department of Anesthesiology's clinical data base[12] were performed. Having approximately 5,000 options in the surgical CPT coding gave a finer subdivision of the cases than did the in-house system.

Data collection forms and instructions were provided to the participating surgeons' patient care coordinators or to the surgeons themselves. The surgeons were asked to complete a data collection form at the time of booking each case, to provide a time estimate, a description of the operation and its CPT code, an estimate of case difficulty (by quintiles), their certainty that the proposed procedure would be performed (or changed due to intraoperative findings), and the presence of confounding factors (not using historical data on operating times for that procedure, lack of essential diagnostic information at the time of booking, or unknown level of resident assistance).

Summary statistics of the phase 2 data were compared with the phase 1 data. A logspline density estimate[13] for the error in the surgeon estimate using the 1994 prospective data was computed and compared with the density estimate for the 1993 retrospective data. In density estimation, it is assumed that observations come from a smooth continuous density function, which is to be estimated from the data (*i.e.*, a smoothed version of a histogram). The logspline method uses polynomial splines to approximate the density. An advantage of the logspline method is that the smoothing parameter, comparable with the number of bins in a histogram, is determined by the algorithm, and does not need to be determined by the user.

In phase 2, regression models were derived to predict the duration of the surgery based on a combination of the surgeon estimate, the estimate provided by the Surgiserver scheduling system, and the other variables collected: gender of the patient, whether the surgery was going to be bilateral, whether the rank of the resident assistant involved was unknown, whether essential diagnostics were missing when the procedure was scheduled, an estimate of the case difficulty (in quintiles), and whether the historical information was used when making the surgeon's time estimate. All linear combinations of these predictors were considered in the models.

L1 regression,[14] a robust regression technique that minimizes the sum of the absolute residuals, was used, rather than the more common least squares regression (which minimizes the sum of the squared residuals). Least squares regression is useful when errors are normally distributed, but it is known to be influenced strongly by extreme outliers. Some of our cases had huge residuals, presumably because the procedure that eventually was carried out differed considerably from the planned procedure or because serious intraoperative problems arose. To avoid biasing the results with those extreme outliers, the L1 technique was chosen. Although most of the available robust regression techniques give similar results, L1 regression has the advantage that its loss function, the sum of the absolute errors, has a natural interpretation as the sum of the time that cases extended beyond their estimates plus the sum of the time that duration of cases was less than their estimates.

Because models with different numbers of predictors were considered, standard measures of goodness-of-fit, such as the coefficient of determination R^2, cannot be used. (The R^2 will increase whenever additional predictors enter a model, even if those do not actually improve the prediction.) Instead, n-fold (leave-one-out) cross validation was used to compare the predictive performance of the models considered. Cross validation involves estimating the parameters in the model from a portion of the data, after which the cases that were not used to estimate those parameters are predicted. Because the predicted cases were not used to estimate the parameters, their prediction errors form an unbiased estimate of the predictive performance of the model.[14]# In n-fold cross-validation, this technique is taken to the extreme, because all cases except one are used to estimate the parameters, and only the one remaining case is predicted. The prediction error for that case is then estimated by comparing the prediction of this model with the actual value. The PRESS[15,16] (predicted residual sum of squares) estimate for the prediction error is obtained by averaging the prediction errors for all cases, so that eventually, all cases are used n-1 times to estimate parameters and once to estimate a prediction error. Using n-fold cross-validation avoids the illusion that a model with more predictors always seems to fit better than a model with a subset of those predictors. Specific models for each specialty and a combined model applied to the complete dataset were derived as described. Methods to calculate standard errors for regression coefficients in robust regression techniques are not as well established as for least squares regression. They were computed using the methods of Sposito[17] and Bassett and Koenker.[18] The required density estimate at the median of the error distribution was obtained using the logspline procedure.

Under the null hypothesis that the model estimates are not better than the scheduling system or surgeons' estimates, the number of cases for which the model estimate is better than the scheduling system or surgeons' estimates has a binomial distribution, with n equal to the number of cases and $P = 0.5$. Because the number of samples is large, estimates may be compared using normal tables.

Phase 3: Prospective Data Collection by Personal Interview

Five surgeons from different specialties (neuro-, oral, general, orthopedic, and ophthalmic surgery) were asked to provide scheduling information by personal interviews for a 3-month period, to assess the effects of improving the fidelity of data capture. For each of these surgeons, the phase 2 (combined) model that was fitted to the 1994 data was applied. For this model, we computed again the median absolute prediction error. In addition, for each of the five surgeons, a separate (personal) model was fitted, using the 1994 and 1995 data. (For several of the surgeons, using just the 1995 data would have left an insufficient number of cases to formulate a reliable model.) As in phase 2, this was an L1 regression model, and an unbiased estimate of the prediction error was obtained using n-fold (leave-one-out) cross-validation.

Results

Phase 1

Of the 8,897 elective cases, 5,667 were analyzable (*i.e.*, had both surgeon and scheduling system estimates available). For all cases combined, the mean absolute error of the surgeons' estimates was 53 min (34% of average case length of 157 min), compared with a mean absolute error of 55 min for the scheduling system. For 26 of the 43 surgeons, the surgeon's score was significantly better ($P < 0.01$) than the scheduling system's score for that same surgeon, based on the binning procedure, whereas no surgeon's estimates were significantly worse than the scheduler at the level of $P = 0.05$. In retrospect, the average of the two predictions would have yielded a smaller mean absolute error of 49.4 min. Choosing retrospectively the better of the two predictions still gave a mean absolute error of 38.5 min.

Six surgeons, from the departments of ophthalmic (2), neuro-, orthopedic, otolaryngology, and oral surgery, performing 10.3% of the cases, had scores below the $P = 0.01$ level when compared with all the surgeons, suggesting they were extremely good predictors. Conversely, five individuals, from the departments of general (2), urologic (2) and thoracic surgery, performing 10.6% of the cases, were extremely poor predictors ($P > 0.99$). If all the surgeons were actually equally good at prediction, only 1% of them would be expected to be good and 1% as bad by their scores. For the six best predictors, the mean absolute error was 35 min (24% of 145 min), and for the five worst predictors, the mean absolute error was 73 min (44% of 166 min). Three of the six surgeons who provided the best predictions were the ones for whom the scheduling system's estimates were the most accurate, whereas three of the five worst predictors were among the four surgeons who had the worst scheduler-generated predictions.

For purposes of comparison with previously published data,[1] table 1 shows the distribution of "accuracy ratios," calculated as the ratio of actual to scheduled hours (from the scheduling system) for all elective cases (including those from surgeons with fewer than 50 cases per year).

Phase 2

Summary statistics of phase 2 are shown in table 2. Approximately one fifth of the collected forms had to be discarded because they were incomplete, were completed after the time of scheduling the surgery, or had been completed for operations scheduled but not done. Of the 711 acceptable forms, a further 5 cases attributed to the discipline "other" were dropped, leaving 706 cases for analysis. The field recording degree of difficulty in quintiles was completed patchily (*e.g.*, only approximately 10% of general surgery records had this field completed). Overall, approximately 30% of cases were assigned to each of the three least difficult

Table 1. Ratios of Actual Case Time to Scheduled Case Time

			Scheduled Hours			Actual Hours		
Ratio	No. of Cases	%	Total	%	Average	Total	%	Average
<0.70	3,282	37	9,190	35	2.8	4,570	19	1.39
0.70–0.85	1,429	16	4,968	19	3.48	3,841	16	2.69
0.85–1.00	1,240	14	4,515	17	3.64	4,141	17	3.34
1.00–1.15	857	10	3,103	12	3.62	3,313	14	3.87
1.15–1.30	528	6	1,822	7	3.45	2,217	9	4.20
>1.30	1,561	18	2,703	10	1.73	5,938	25	3.80
Total	8,897	100	26,301	100	2.96	24,021	100	2.70

Actual case time: incision to application of dressing.

quintiles, and only 10% to the most difficult two quintiles combined. All specialties reported a greater than 90% confidence that the scheduled operation would be done as scheduled, with the exception of orthopedic surgery (67% confidence). Historical data were used in estimation of 81% of cases by all specialties except ophthalmology (18%). All specialties reported a less than 20% incidence of cases where essential diagnostic information was missing at the time of booking. There was wide variation in knowledge of resident assistance.

Figure 1 shows the logspline fitted density functions of the error in surgeon estimates for phase 2 and phase 1 data. It can be seen that the surgeons in phase 2 further underestimated the duration of surgery, because the density estimate is centered further away from zero. Otherwise, the density of prediction errors seems to be similar to that of phase 1 (30.5% of the average case length of 168 min).

Models derived using combinations of the possible confounding factors (such as lack of historical data, essential diagnostic data missing, and difficulty of the case) yielded prediction errors that were, at best, 5% smaller than the model that depended only on the surgeon and scheduling system estimates. Using only those parameters, specialty specific and combined models using the complete dataset were derived (table 2). It can be seen that the models for gynecology, general surgery, ophthalmology, and oral surgery are essentially the same as the combined model, giving approximately 25% more weight to the surgeons' estimates than to the scheduler's estimates. For urology, the model gives more weight to the scheduling system estimate, whereas for orthopedics, it gives three times the weight to the surgeon estimate. The model for otolaryngology, having the smallest sample size, is much more variable than the other models.

There is a small improvement of the specialty specific model estimates over both the scheduling system estimates and the surgeon estimates (table 2), except for

Table 2. Summary Statistics and Model Prediction Results—Phase 2 Data

Discipline	n	MOT	SE	SSE	SME	CME	Estimate (min)	St Error
Gynecology	88	186	33.5	34.2	31.6	29.7	−6.52 + 0.67*SE + 0.53*SSE	0.07–0.13
Gen Surg.	134	252	29.4	27.5	24.4	24.5	4.95 + 0.68*SE + 0.40*SSE	0.07–0.13
Otolaryngology	30	208	35.1	34.4	31.3	31.1	−25.8 + 0.73*SE + 0.71*SSE	0.4
Urology	129	146	34.0	32.4	31.5	30.9	−2.99 + 0.41*SE + 0.66*SSE	0.055
Ophthalmology	218	129	30.8	31.2	27.5	26.9	−0.09 + 0.69*SE + 0.46*SSE	0.07–0.13
Oral	50	131	28.9	32.8	31.2	28.2	−7.58 + 0.68*SE + 0.41*SSE	0.17
Orthopedic	57	155	19.1	21.5	19.9	16.9	−3.85 + 0.92*SE + 0.29*SSE	0.07–0.13
Combined	706	168	30.5	30.3	27.7	26.8	−0.51 + 0.62*SE + 0.49*SSE	0.035

Because of the various approximations made in calculating the standard errors (see text), they should be treated as suggestive.

MOT = mean operating time (minutes); SE = mean error, surgeons' estimate (% of MOT); SSE = mean error, scheduling system estimate (% of MOT); SME = mean error, specialty specific model (% of MOT); CME = mean error, combined model (% of MOT); St Error = approximate standard error of the coefficients of SE and SSE.

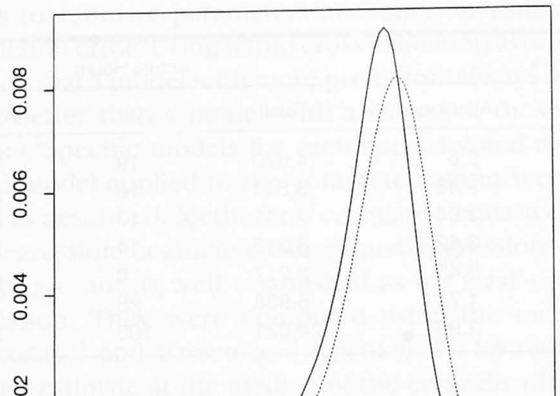

Fig. 1. Logspline estimate of the density function of the error in surgeon estimates for phase 1 (solid line) and phase 2 (dotted line). The phase 2 curve is shifted rightward, which indicates larger mean prediction errors, whereas the variability of the prediction errors is essentially unchanged.

the prediction errors for the separate models for oral and orthopedic surgery, which yield errors that are slightly worse than the surgeon estimates. The combined model improves 11.5% over the scheduling system estimate and 12.1% over the surgeons' estimate, whereas the specialty specific model improves only 8.6% and 9.1%, respectively. The combined model estimate was better than the scheduling system estimate in 421 of 706 cases ($P < 0.0001$), and better than the surgeons estimate in 424 of 706 cases ($P < 0.0001$).

Phase 3

The phase 3 results are summarized in table 3. The overall errors in estimation were similar to those of phase 2 (30.5% of average case length of 225 min). For the general surgeon (for whom the models yielded the largest improvements), inclusion of the surgeons' estimates (in quintiles) about difficulty of the case further reduced the mean absolute prediction error from 17.8% to 16.9% of the average case length. For none of the other surgeons was there such an improvement. For the five surgeons combined, applying the personal models gave an improvement of 17.2%, whereas the model from phase 2 yielded an improvement of 18.2% over the scheduling system estimate (compared with an improvement of 11.5% in phase 2). However, without the general surgeon, these improvements are down to 11.0% for the combined model, whereas the personal model predicts 6.0% worse than the scheduling system predictions.

Discussion

Whatever system is used to schedule operating room time,[7-11] having more accurate estimates of each case's duration should help reduce both underutilization and overrunning of the planned workday. The Surgiserver's trimmed mean (eight of ten) estimation method used in our institution gave results that were statistically no better than any individual surgeon's estimates and were worse than the estimates of 60% of the surgeons. Therefore, this software offers no advantage over simply asking the surgeons to provide a time estimate when they book each case, although the accuracy of their estimates is also low. Trimmed mean estimation is widely used in operating room scheduling, because it is provided by both Surgiserver, which has approximately a one-third share of the proprietary scheduling systems market, and the competing Orbit Surgical Services Management software (Enterprise Systems, Wheeling, IL), which also has approximately a one-third share of the market (personal communications, Ann Randall, Serving Software, and Tom Newell, Enterprise Systems). It was beyond the scope of our study to investigate the validity or otherwise of this approach; our intent was to assess the accuracy of surgeons' estimates compared with a widely used system, and to investigate whether computer models that include surgeon-provided data could improve its accuracy.

There are several possible explanations for the inaccuracy of estimates generated from the scheduling sys-

Table 3. Model Prediction Results—Phase 3 Data

Discipline	n	MOT	SE	SSE	PME	CME
Neurosurgeon	23	332	34.3	28.8	37.4	26.5
Oral surgeon	10	135	28.4	38.4	37.1	28.7
Gen. surgeon	43	270	29.7	33.6	17.8	24.4
Ortho. surgeon	32	178	27.0	23.3	21.1	22.1
Ophth. surgeon	15	87	34.4	38.4	35.4	34.1
All	123	225	30.5	30.5	25.6	25.0

MOT = mean operating time (minutes); SE = mean error, surgeons' estimate (% of MOT); SSE = mean error, scheduling system estimate (% of MOT); PME = mean error, personal model (% of MOT); CME = mean error, combined model (% of MOT).

tem's historical database. The in-house coding scheme for classifying the operations lacks the specificity of CPT coding, and, therefore, lumps together procedures of differing complexity and time requirements. However, using the more specific CPT coding scheme would have required a larger data base and a longer data collection period to initialize the prediction algorithm. In addition, among identical operations, potentially confounding variables that are unknown to the scheduling system, such as obesity, adhesions from prior surgery, difficult anatomy, *etc.*, may have a major impact on the operating time for particular patients. Because such variables tend to be specialty specific, it is difficult to conceive of a data collection form that would be sufficiently comprehensive to allow the scheduling system to account for such variables without being unduly cumbersome to complete and process. We hypothesized that surgeons would be able to take individual patient and specialty specific factors into account to provide more accurate estimates.

We chose a different definition of operating time (from incision to dressing application) in the study than the Association of Anesthesia Clinical Directors' (patient entering to leaving the operating room). This decision was made because that definition had been used in the collection of the retrospective data, and because our intent was to address how well surgeons could predict the duration of surgery alone (*i.e.*, the part of the schedule they control). Therefore, we assumed that the majority of the variability in the schedule was accounted for by discrepancies in the length of actual *versus* scheduled operating time. This assumption is supported by Dexter *et al.*,[19] who noted that even if the anesthesia-controlled portion of the time (room entry to positioning time plus wound dressing to room exit time) could be decreased to zero, insufficient extra time would be made available, on average, to allow one additional 30-min operation to be performed.

Despite their sophistication and extensive information base, the surgeons in phase 1 were only slightly better at estimating than the scheduling system, and they tended to underestimate their operating time. Possible reasons for the poor surgeon predictions were: 1) inattention to the prediction process because of knowledge that their input was not being used, 2) lack of historical information for them to base their estimates on, and 3) a tendency to give the same time estimate for every operation of a given type, much like the scheduling system did. The phase 2 estimates by the surgeons were less accurate than the phase 1 estimates, despite their being asked to be conscientious about the process of estimating.

Operations by some surgical specialties may be inherently easy or inherently difficult to predict. Gordon *et al.*[1] noted, for example, that at The Johns Hopkins Medical Institutions, laparoscopies (the second-most-frequent procedure) varied by 42% from the estimated time, whereas hysterectomies, inguinal hernia repairs, and prostatectomies varied by only 0-4%. Our data support the contention that some types of surgery are inherently difficult to predict: three of the six best surgeon predictors also were those for whom the scheduling system estimates were most accurate and three of the worst were among the four surgeons who had the worst scheduling system-generated predictions. The good predictors were from specialties that operate on the body surface or extremities, where diagnosis is usually straightforward and the operations are standardized. By contrast, the poor predictors were from specialties that, in this institution, have a high proportion of intracavitary, oncologic procedures, in which the surgical procedures are far from standardized and may have to be modified during the course of the operation.

Personal interviews with the surgeons identified as being good predictors did not help elucidate the reasons for their success, although a common theme was that they took the process of prediction seriously and made efforts to be accurate even though that accuracy was not rewarded by, for example, greater availability of operating time. Indeed, in some hospitals, accurate prediction might penalize a surgeon trying to "fit" an extra case into his or her elective scheduling block. The fact that, even retrospectively, the reduction in the error is quite small, suggests that some modeling is required to reduce the prediction errors by an amount that would be practically important.

The second phase of the study was designed to provide historical data to the surgeons and to collect potential explanatory variables, unavailable to us in the retrospective data, for use in predictive modeling. We attempted to identify possible confounding factors, common to all specialties, that would allow us to derive more accurate models. The results of this approach were disappointing, in that the more complex models produced results no better than those from simpler models using only surgeon and scheduling system estimates. It is possible that different confounding factors would have been more useful in this regard. We believe, however, that this merely confirms our original supposition that surgeons' evaluation of patients is a sophisticated process that incorporates a large number of specialty-specific factors that affect the duration of an operation. If this is true, modeling using historical data (*i.e.*, the scheduling system estimates) to compensate for systematic biases in the surgeons' estimations may be the best approach. Accordingly, we derived models using only surgeon and scheduling system estimates.

The model estimates showed a modest improvement in accuracy, but the absolute errors were still of a magnitude that would make accurate scheduling difficult overall. However, even small percentage improvements may produce large real-time savings in longer or high volume cases (see later). Calculating specialty-specific models did not improve on the combined model, even though specific models weighted the surgeon and scheduling system estimates differently. In addition, only the otolaryngology model (having the smallest sample size) was markedly more variable than either the combined model or the other specialty-specific models. That the models gave results similar both absolutely and in terms of variability suggests that the approach of using a combined model to predict individual specialties is valid, although it is possible that larger specialty-specific datasets would improve the accuracy of separate models. A sophisticated information system would be able to "learn" from experience and adjust models to take this into account; as numbers in individual datasets increased, it would be possible to choose the more accurate model (combined or specialty specific).

The lack of improvement in the surgeon's performance in phase 2 could have been due to several factors. The low rate of return of data forms (approximately 15% of the possible cases) and the fact that many were sloppily prepared suggest a lack of commitment to the study by the surgeons. All surgeons who performed more than one case per week were approached, and gave consent to be part of the study. The surgeons who actually participated showed a certain level of interest, in that they made time to complete and return the forms. It is speculative as to the influence the specialties not participating would have had on the results. Cardiac surgery has been shown to be predicted accurately at a different institution, actual time varying only 12% from scheduled.[1] The neurosurgeon who agreed to participate in phase 3 at UWMC had a percentage prediction error of 34.3% (compared with 30.5% overall).

Despite extensive educational efforts, both written and in person, it was uncertain whether the instructions for form completion were clear enough. For example, the distribution of the quintiles of difficulty tended to be biased to the middle (averagely difficult) quintile and were not uniformly distributed, as true quintiles would have been. In the great majority of cases, the surgeons indicated on the data collection form that they were confident or fairly confident of their estimates, and in only a small minority of cases was diagnostic information lacking. In other words, the surgeons thought, in the majority of cases, that the estimates they were giving were as accurate as possible.

The third phase was designed to overcome the difficulties with surgeon compliance in the second phase. It was decided to concentrate on a small number of interested surgeons with varying lengths and types of operations, in an effort to obtain complete data for their bookings and to further encourage conscientious efforts at time estimation. Each was approached before every elective case during a 3-month period by the same research coordinator, and a form identical to that used in phase 2 was completed. Results were encouraging—personal modeling reduced the error in surgeons' estimates by 18% in phase 3, compared with 11.7% in phase 2. The absolute errors were still large, however, and if the best individual surgeon predictor is removed, the improved accuracy from modeling is comparable with that in phase 2. However, for that predictor, a busy general surgeon with a heterogeneous caseload and a major commitment to intraoperative teaching, personal modeling reduced the prediction error by 40%. His personal model error (17.8%) compares favorably with the errors of even the six best predictors from phase 1 (24%). In absolute terms, his personal model produces estimates that are more accurate than the scheduling system by 43 min. Assuming he does two cases daily, that provides potential error reduction of 86 min, sufficient time to fit in an extra case (if the present scheduler were overestimating) or suggest postponing one case (if the present scheduler were underestimating).

Personal modeling increased the prediction error for three of the five surgeons in phase 3, which could be caused by large variance due to small numbers in the dataset. This explanation is supported by the data (table 3), which show: 1) the largest error reductions (between the surgeon and personal model estimates) occurred with the larger sample sizes, and 2) a reduction in prediction error was seen for every surgeon in the combined model, which was derived from a greater number of cases. We can only speculate as to whether having increased numbers of cases for individual surgeons would make their personal models generally better than the combined model.

Relation to Earlier Work

In the largest reported study of operating times, Gordon *et al.*[1] compared surgeons' estimates with actual case duration for 56,000 cases throughout a 3-yr period at Johns Hopkins Hospital, an academic medical center very similar to ours. Scheduling was entirely based on surgeons' estimates, and no attempts were made to compare scheduling system-generated estimates or to improve accuracy with computer modeling. The average case length at Johns Hopkins was slightly shorter than ours, but the variability of their surgeons' errors of estimation were similar. Only 26% of their case lengths were within 15% of the estimate (*vs.* 24% at UWMC in 1993; table 1), whereas 31% of cases were more than 30% shorter than the estimate (37% at UWMC), and 19% were more than 30% longer than estimated (18% at UWMC).

At UWMC, there was no relation between accuracy of prediction and scheduled case duration: cases that over- or underran were predicted to last, on average, approximately the same. Cases scheduled to be shorter appear to be clustered at the extremes (ratios <0.7 or >1.3), owing to similar absolute errors representing larger percentage errors on shorter scheduled cases compared with longer scheduled cases. The apparent increase in average length with increasing ratio of actual to scheduled case time in our data and those from Johns Hopkins is artifactual, and would be similar even if all cases had been estimated to last the same time. Gordon *et al.*[1] emphasize that, when trying to improve accuracy of scheduling, focusing on high volume procedures and lengthy procedures can yield gains in productivity. Even though such cases are scheduled no more inaccurately than any others, it is self-evident that small percentage changes in these areas would produce large real-time benefits.

Overall, the reported 2% difference between scheduled and actual operating time at Johns Hopkins (*vs.* 9.5% at UWMC) appears, at first glance, to represent a good match between workload and capacity. However, these figures conceal the fact that underuse and overruns on individual days will both increase costs while canceling each other in the utilization calculation. In addition, within a working day, noninterchangeable operating rooms, equipment, and personnel may have prevented cases from being moved from rooms running late to those finishing early, although the overall utilization figure could still approach 100%. Overall utilization figures also conceal inefficiencies resulting from deviations from scheduled arrival and departure times on other departments dependent on the operating room, such as holding and recovery areas,[5] radiology suites, *etc.* Although it would be informative to examine the variability in the length of the case lists for individual rooms, those data are not available from the Johns Hopkins study or from our own work.

In another study that actually used statistics to help generate daily operating lists, Rose and Davies[20] gathered data on 612 urologic operations performed during a 15-month period in one hospital. They arranged their data in aggregates of short (3.5-h) and long (7-h) operating sessions. Of the 49 types of operations, 15 occupied 86% of the operating time. For those 15 types of operations, the "loading standard" (a weighted average based on the statistical beta distribution and conceptually borrowed from critical path analysis used in industry) was calculated and supplied to the surgeons to aid in booking their cases during the second phase of the study. They demonstrated a significant reduction in the standard deviation of the length of the operating sessions (from 71 to 35 min for the short block and from 99 to 77 min for the long block), while keeping the mean length of the operating sessions (and, therefore, operating room productivity) unchanged. Although encouraging as regards our own work, their study was limited, because it examined only one surgical specialty, whose cases are, mostly, relatively homogeneous. Transurethral surgery, for example, would seem to be inherently predictable, because the time taken depends on the weight of the enlarged gland alone. This contention is supported by the Johns Hopkins data,[1] in which transurethral operations have an error of only 15% from scheduled estimates.

Although not directly comparable with our study, previous work has shown the importance of concentrating on surgical time, rather than anesthetic or room turnover time. Mazzei[21] noted that attempts to shorten room preparation, induction and wake-up, room cleanup, and surgical preparation and draping times would require special effort, more manpower, and capital expenditure. He noted that in an academic medical center, however, such efforts would save only approximately 30 min daily, which would not provide enough additional time to perform more operations. Dexter *et al.*[19] made the similar point that decreasing anesthesia controlled time by using, for example, preoperative intravenous catheter teams, procedure rooms, or shorter-acting drugs may simply increase costs, without providing commensurate savings in useful operating room time. By contrast, more accurate scheduling of cases requires only redeployment of existing data collection personnel and computer facilities, because most institutions use some form of computer-aided scheduling already. It is beyond the scope of this study to investigate the economic impact of improved accuracy of scheduling, but the potential benefits of matching demand with resources for a facility as expensive as an operating room suite, and the ancillary services that depend on it, are obvious.

Limitations of the Current Study

The scope and sample size of this study were constrained severely by the available resources and the state of information management within the UWMC. Ideally, we would have had available a comprehensive, hospital-wide information system to integrate our study protocol into the active scheduling process. Thereby, a surgeon wishing to schedule a case could have been given timely historical data on operating times for similar cases, as well as feedback on the accuracy of previous predictions. In addition, viewing of the historical data and completion of our questionnaire could have been made mandatory to book a case through the scheduling system, allowing us to acquire a complete set of data. Validity checking of surgeon-supplied data could have been performed on-line, and statistical tests could have been used to identify and discourage some surgeons' practice of always estimating the same time or quintile of difficulty for a particular type of case. Finally, performing the study for a longer time may have improved the predictive models because, for example, enough data would be accumulated to make the quintiles more specific for the types of operations. Then, in a production scheduling system, the statistical model building could have been fine-tuned and made adaptive to changing conditions.

The general applicability of these results is limited, because the study was performed within a tertiary referral center. Our average case length of 157 min is considerably longer than the average expected in community hospitals. Having fewer surgeons with greater individual workloads, more standardized patients, less complex operations, and no resident teaching would undoubtedly have resulted in better predictions overall.

Conclusions and Suggestions for Further Work

Encouragingly, as a whole, surgeons were better than the commercial scheduling software at estimating the duration of surgery, and individual surgeons did much better. A simple model that combined the surgeon's estimate with the historical data reduced the prediction errors significantly. The other variables examined in this study did not improve on that model. However, we would not conclude that such covariates are useless until they have been tried on a larger data set, with conscientious surgeon effort and timely feedback. Future studies are worthwhile, therefore, to determine: 1) whether improvement would result by integrating the data gathering and analysis into the surgical booking software as envisaged; 2) whether better results would be obtained in a community hospital setting; and 3) whether the improvements in prediction will translate into more efficient scheduling and lower costs. Surgeons also need an incentive to reduce their errors in estimating duration. As the impact of managed care grows, the incentive to be more accurate will become greater.

The authors thank the surgeons who generously gave of their time to participate in this study, Katherine Cox and Sonja Kapitan, for assisting with the data collection, and Greta Marcelja and Judi Pence, for help in getting access to the computer databases.

References

1. Gordon T, Paul S, Lyles A, Fountain J: Surgical unit time utilization review: Resource utilization and management implications. J Med Syst 1988; 12:169-79

2. McQuarrie DG: Limits to efficient operating room scheduling. Lessons from computer-use models. Arch Surg 1981; 116:1065-71

3. Gil AV, Galarza MT, Guerrero R, de Velez GP, Peterson OL, Bloom BL: Surgeons and operating rooms: Underutilized resources. Am J Public Health 1983; 73:1361-5

4. Barr A, McNeilly RH, Rogers S: Use of operating theatres. BMJ 1982; 285:1059-61

5. Dexter F, Tinker JH: Analysis of strategies to decrease postanesthesia care unit costs. ANESTHESIOLOGY 1995; 82:94-101

6. French S: Sequencing and scheduling: An introduction to the mathematics of the job-shop. New York, Wiley, 1982

7. Magerlein JM, Martin JB: Surgical demand scheduling: A review. Health Serv Res 1978; 13:418-33

8. Przasnyski ZH: Operating room scheduling. A literature review. AORN J 1986; 44:67-79

9. Hamilton DM, Breslawski S: Operating room scheduling. Factors to consider. AORN J 1994; 59:665-8, 671-80

10. Hancock WM, Isken MW: Patient-scheduling methodologies. J Soc Health Syst 1992; 3:83-94

11. Hancock WM, Walter PF, More RA, Glick ND: Operating room scheduling data base analysis for scheduling. J Med Syst 1988; 12:397-409

12. Bashein G, Barna CR: A comprehensive computer system for anesthetic record retrieval. Anesth Analg 1985; 64:425-31

13. Kooperberg C, Stone CJ: Logspline density estimation for censored data. J Comput Graph Stat 1992; 1:301-28

14. Huber PJ: Robust statistics. New York, Wiley, 1981

15. Allen DM: The relation between variable selection and prediction. Technometrics 1974; 16:125-27

16. Stone M: Cross-validatory choice and assessment of statistical predictors. J R Stat Soc, Series B. 1974; 36:111-47

17. Sposito VA: Some properties of Lp estimators, Robust Regression. Edited by Lawrence KD, Arthur JL. New York, Dekker, 1990, pp 23-58

18. Bassett G Jr, Koenker R: Asymptotic theory of least absolute error regression. J Am Stat Assoc 1978; 73:618-622

19. Dexter F, Coffin S, Tinker JH: Decreases in anesthesia-controlled time cannot permit one additional surgical operation to be reliably scheduled during the workday. Anesth Analg 1995; 81: 1263-8

20. Rose MB, Davies DC: Scheduling in the operating theatre. Ann R Coll Surg Engl 1984; 66:372-4

21. Mazzei WJ: Operating room start times and turnover times in a university hospital. J Clin Anesth 1994; 6:405-8